The support vector classifier (1)

Manel Martínez-Ramón

ECE, UNM

- The Support Vector Machine idea consists of minimizing the previous empirical risk plus the structural risk through margin maximization, this is:

$$
\begin{aligned}
\text{minimize } & L_p(\mathbf{w}, \xi_n) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n \\
\text{subject to } & \left\{ \begin{array}{c} y_n\left(\mathbf{w}^\top \mathbf{x}_n + b\right) > 1 - \xi_n \\ \xi_n \geq 0 \end{array} \right.
\end{aligned}
$$

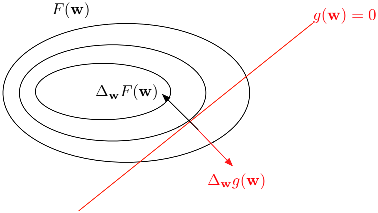
- C is a free *tradeoff* parameter.
- Subindex p stands for *primal*. We'll have a *dual* later.

Lagrange optimization

- In order to optimize the machine we need some Lagrange minimization.
- Assume the following minimization with constraints

$$\text{minimize } F(\mathbf{w})$$
$$\text{subject to } g(\mathbf{w}) = 0$$

- The optimal point is clearly where both gradients are proportional.



- Roughly speaking, we must then construct the functional

$$L_{Lagrange} = F(\mathbf{w}) - \alpha g(\mathbf{w})$$

where $\alpha \geq 0$ is a Lagrange multiplier or *dual* variable.

- The optimization consists of computing the gradient wrt the primal variables $\mathbf{w}$ and nulling it.

$$\Delta_{\mathbf{w}} F(\mathbf{w}) - \alpha g(\mathbf{w}) = 0$$

This will lead to the Karush Kuhn Tucker (KKT) conditions.

- Then, we must find the value of the dual variables.

Optimization of the SVC

- The SVM primal problem is

$$\text{minimize } L_p(\mathbf{w}, \xi_n) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{n=1}^{N}\xi_n$$

$$\text{subject to } \begin{cases} y_n\left(\mathbf{w}^\top \mathbf{x}_n + b\right) - 1 + \xi_n \geq 0 \\ \xi_n \geq 0 \end{cases}$$

- We must use Lagrange multipliers to change the constrained problem into an unconstrained one.
- Since there are $2N$ constraints, we need $2N$ multipliers, namely α_n for the first set, and μ_n for the second one

- The Lagrangian is then

$$
\begin{aligned}
L_L(\mathbf{w}, \xi_n, \alpha_n, \mu_n) =& \frac{1}{2}||\mathbf{w}||^2 + C \sum_{n=1}^{N} \xi_n \\
& - \sum_{n=1}^{N} \alpha_n \left(y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) - 1 + \xi_n \right) \\
& - \sum_{n=1}^{N} \mu_n \xi_n
\end{aligned}
$$

subject to $\alpha_n, \mu_n \geq 0$, and where the primal variables are $\mathbf{w}$ and ξ_n.

- We first null the gradient with respect to $\mathbf{w}$.

$$\Delta_{\mathbf{w}} L_L(\mathbf{w}, \xi_n, \alpha_n, \mu_n) = \mathbf{w} - \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n = 0$$

- This give us the result

$$\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n$$

- or, in matrix notation

$$\mathbf{w} = \mathbf{X}\mathbf{Y}\boldsymbol{\alpha}^{\top}$$

where $\mathbf{Y}$ is a diagonal matrix containing all the labels and $\boldsymbol{\alpha}$ contains all the multipliers.

The KKT conditions

- Then we null the derivative wrt the slack variables ξ_n and b.

$$\frac{\partial}{\partial \xi_n} L_p(\mathbf{w}, \xi_n, \alpha_n, \mu_n) = C - \alpha_n - \mu_n = 0$$

$$\frac{d}{db} L_p(\mathbf{w}, \xi_n, \alpha_n, \mu_n) = -\sum_{n=1}^{N} \alpha_n y_n = 0$$

- Also, we must force the complementarity property over the constraints

$$\mu_n \xi_n = 0$$

$$\alpha_n \left(y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) - 1 + \xi_n \right) = 0$$

The KKT conditions

- In summary, the KKT conditions are

$$
\begin{aligned}
\mathbf{w} &= \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n && (1)\\
C - \alpha_n - \mu_n &= 0 && (2)\\
\sum_{n=1}^{N} \alpha_n y_n &= 0 && (3)\\
\mu_n \xi_n &= 0 && (4)\\
\alpha_n \left(y_n \left(\mathbf{w}^\top \mathbf{x}_n + b \right) - 1 + \xi_n \right) &= 0 && (5)\\
\alpha_n \geq 0,\ \mu_n \geq 0,\ \xi_n &\geq 0 && (6)
\end{aligned}
$$

The solution

- From (2) and (4)

$$
\begin{aligned}
C - \alpha_n - \mu_n &= 0 \\
\mu_n \xi_n &= 0
\end{aligned}
$$

  we see that if $\xi_n > 0$ (sample inside the margin or misclasified), then $\alpha_n = C$.
- With (5), we see that if the sample is on the margin, $0 < \alpha_n < C$
- If the sample is well classified and outside the margin, then $\xi_n = 0$, and (5) determines that $\alpha_n = 0$.

- The estimator $y_k = \mathbf{w}^\top \mathbf{x}_k + b$ can be rewritten by virtue of (1) as

$$y_k = \sum_{n=1}^{N} y_n \alpha_n \mathbf{x}_n^\top \mathbf{x}_k + b$$

or, in matrix notation

$$y_k = \boldsymbol{\alpha}^\top \mathbf{Y}\mathbf{X}^\top \mathbf{x}_k + b$$

Outcomes of the lesson

- From the primal expression of the SVM functional, we have constructed a Lagrange functional
- By computing the derivatives of the Lagrangina wrt the primal parameters, we have found:
  - The support vectors
  - A dual expression of the classifier as a function of them.